文章编号: 1673-3193(2025)05-0611-11 J. North Univ. China, Nat. Sci. Ed., 2025,46(5):611-621.

基于关系标签语义与全局特征融合的专利实体关系抽取

张斌龙[1], 雷海卫[1], 李成奇[2], 智 媛[1]

(1. 中北大学 计算机科学与技术学院, 山西 太原 030051;
2. 陕西飞机工业有限责任公司, 陕西 汉中 723200)

摘 要: 专利文本长度较长且关系重叠严重, 传统的关系抽取模型中将关系视作离散的标签且只关注文本的局部特征, 未充分利用其中的潜在信息, 影响复杂的专利文本的抽取性能。针对以上问题, 本文在OneRel模型的基础上提出了一种基于关系标签语义与全局特征融合的填表式实体关系抽取模型SRL-GFI(Semantics of Relation Labels and Global Feature Integration)。SRL-GFI模型使用BERT预训练模型获取关系标签的语义信息, 并与表格标记方式相结合得到关系-标记联合编码, 使之与分词对的编码进行匹配得出相应的相关性分数, 再通过全局信息融合模块将模型中间结果反馈回去并迭代, 以融合其中的全局特征, 从而进行实体关系抽取。多个数据集上的对比实验都显示了新模型性能的提升, 其中在专利实体抽取数据集PERD和TFH2020上, $F1$值分别达到79.8%和55.9%, 相比OneRel提升了5.0百分点和1.8百分点。本文提出的SRL-GFI模型有效利用了关系标签中的语义信息, 以及文本中的全局信息和关系三元组中的内在联系, 可以很好地应用于专利领域中复杂文本的实体关系抽取。

关键词: 实体关系抽取; 关系标签语义; 全局特征融合; 表格填充; 专利领域

中图分类号: TP391.1 **文献标识码:** A **doi:** 10.62756/jnuc.issn.1673-3193.2024.11.0012

引用格式: 张斌龙, 雷海卫, 李成奇, 等. 基于关系标签语义与全局特征融合的专利实体关系抽取[J]. 中北大学学报(自然科学版), 2025, 46(5): 611-621.

ZHANG Binlong, LEI Haiwei, LI Chengqi, et al. Patent entity relation extraction based on the semantics of relation labels and global feature integration[J]. Journal of North University of China(Natural Science Edition), 2025, 46(5): 611-621.

Patent Entity Relation Extraction Based on the Semantics of Relation Labels and Global Feature Integration

ZHANG Binlong[1], LEI Haiwei[1], LI Chengqi[2], ZHI Yuan[1]

(1. School of Computer Science and Technology, North University of China, Taiyuan 030051, China;
2. Shaanxi Aircraft Industry Co., Ltd., Hanzhong 723200, China)

Abstract: Patent texts are lengthy and contain heavily overlapping relationships, while traditional relation extraction models treat relations as discrete labels and focus only on local features of the text, failing to fully utilize the latent information, thereby affecting the extraction performance on complex patent texts. To address these issues, this paper proposed a table-filling entity relation extraction model SRL-GFI (Semantics of Relation Labels and Global Feature Integration), based on the OneRel model, which integrated the semantics

---

收稿日期: 2024-11-21

作者简介: 张斌龙(2000—), 男, 硕士生, 主要从事关系抽取的研究。

通信作者: 雷海卫(1980—), 男, 副教授, 博士, 主要从事自然语言处理等方面的研究。E-mail: lhw0312@nuc.edu.cn。

of relation labels and global features. The SRL-GFI model utilized a BERT pre-trained model to obtain semantic information of relation labels and combined it with tagging scheme to derive a relation-marking joint encoding. This encoding was then matched with embedding of word pairs to obtain the corresponding relevance scores. Subsequently, the global feature integrating module fed back and iterated the intermediate results of the model to integrate global features, thereby performing entity relation extraction. Comparative experiments on multiple datasets show performance improvements of the new model, with $F1$ scores reaching 79.8% and 55.9% on the patent entity extraction datasets PERD and TFH2020, respectively, representing improvements of 5.0 and 1.8 percentage points over OneRel. The SRL-GFI model proposed in this paper effectively utilizes the semantic information in relation labels, as well as global information in the text and intrinsic connections within relation triplets, making it well-suited for entity relation extraction of complex texts in the patent domain.

Key words: entity relation extraction; semantics of relation labels; global feature integration; table filling; patent domain

0 引 言

专利文献作为高价值密度的数据源，包含了全球范围内的大量最新技术信息，具有丰富的题录、详尽的内容和规范的格式[1]。随着近年来专利数量的快速增长，传统的人工查阅方法难以应对其激增的趋势。为了从专利文献中批量获取结构化的数据，需要对专利文本进行深入的数据挖掘工作。关系抽取关注事物之间的相互联系，为进一步的信息抽取和知识图谱构建提供重要支撑。

关系抽取旨在从自然语言文本中提取实体之间的关系，获得有效的语义知识，并以(主体，关系，客体)的形式记录。作为信息抽取[2]与构建知识图谱[3]的上游任务，关系抽取被广泛应用到了各个领域自然语言处理的相关任务中。

关系抽取中普遍存在关系重叠的问题：单实体关系重叠(Single Entity Overlap，SEO)，即一个实体与多个实体存在关系；实体对关系重叠(Entity Pair Overlap，EPO)，即一对实体之间存在多个关系；实体嵌套(Subject Object Overlap，SOO)，即一个实体中包含另一个实体的现象。而在专利文本中，文本长度较长，关系重叠的问题更为严重[4]。如在以下这句话中，“一种紫外线杀菌装置，包括紫外线杀菌灯、连接头和定位板，所述连接头设置在所述紫外线杀菌灯顶部；所述连接头包括基座和卡钩，若干个所述卡钩设置在所述基座的端部，相邻的所述卡钩之间留有间隙；所述定位板设置有卡孔，所述连接头的卡钩可卡入至所述卡孔中”，与“杀菌装置”有“主附件”关系的实体的有“紫外线杀菌灯”“连接头”“定位板”，“连接头”和“卡钩”之间存在的关系有“主附件”和“设置有”。

传统的模型忽略了关系标签的语义信息以及文本的全局特征，导致模型难以充分捕捉文本与关系之间的深层次关联，进而在处理复杂专利文本时性能受限。其中，OneRel[5]方法基于表格填充，并采用HornTagging策略来标记表格，这种策略关注主体与客体在表格中形成区域的左上角、右上角及左下角，并为不同的关系类别列出各自的表格，有效解决了关系重叠的问题，但仍需要进行以下优化：

一方面，OneRel将关系表示建模为线性层，忽略了关系标签本身所蕴含的语义信息，并且与文本的编码存在巨大的异质性，这种异质性使得模型难以有效捕捉语义空间中文本与关系标签间的内在联系。本文利用BERT(Bidirectional Encoder Representations from Transformers)预训练模型挖掘关系标签中预设的语义，相较于一般方法中的线性层建模更能体现关系与文本之间的关联。

另一方面，OneRel模型主要关注文本的局部特征，这使得它在处理跨越较长文本距离的实体之间的语义联系时存在不足。此外，对于复杂文本中的关系三元组间的联系，OneRel模型也未能充分利用文本的全局特征进行挖掘。这些局限性限制了模型在处理复杂文本关系抽取任务时的性能。本文利用关系对之间的依赖，将初步识别到的结果反馈回模型中，增强文本的全局特征以及对文本中的其他三元组的识别，很好地利用了关系之间的联系与句子文本中的全局特征。

针对这个问题，本文在OneRel的基础上做出了改进，提出一种基于关系标签语义与全局特征融合的填表式关系抽取方法，主要贡献包括：1) 使用BERT预训练模型挖掘关系标签的语义信

息，并与表格的标记方式相结合，得到关系-标记联合编码使之与分词对的编码相匹配以计算相关性分数，用以增强对分词对类型的判断。 2）引入全局特征融合模块，将模型的中间结果反馈回句子编码中，以此融合全局特征。 3）在多个数据集上进行的实验表明，SRL-GFI 相较于基准模型 OneRel 在 PERD、TFH2020 专利实体抽取数据集，以及 NYT、CONLL04 和 SCIERC 通用数据集上的 $F1$ 值均有提高，实现了性能提升。

1 相关工作

1.1 流水线及联合抽取方法

基于深度学习的关系抽取方法主要分为流水线与联合抽取方法[6]。早期的关系抽取主要采用流水线的方法，将任务分解为实体识别和关系分类两个独立的子任务[7-8]。这种方法虽然灵活，但容易受到误差传递的影响，且缺乏实体抽取与关系判断两个子任务之间的交互[9]。

为了解决这些问题，后续研究提出了联合抽取方法[10-13]。联合抽取方法不再将关系抽取区分为命名实体识别和关系判断，而是同时进行这两个任务，以端到端的方式进行关系抽取，在统一的模型中进行优化，减少误差传递，增强了实体抽取与关系判断之间的交互性。Wei 等[10]提出了将关系三元组中的主体映射到客体的函数的层叠指针模型 CasRel。Yan 等[11]提出了一种分区过滤网络 PFN，将特征编码分解为划分和过滤两个步骤，对命名实体识别与关系抽取两个任务间的双向交互进行建模，通过共享参数的方法确保任务特定特征的编码彼此相关。Zheng 等[12]提出了一种基于实体抽取与全局对齐的联合抽取方法 PRGC，并使用关系判断模块去除其中的冗余关系矩阵。Wang 等[13]提出了一种握手标记方案的模型 TPLinker，将联合抽取建模为标记对链接的问题，对齐每个关系类型中实体对的边界标记，使用单阶段解码方法，有效缓解了关系重叠的问题。

1.2 基于填表方式的关系抽取

基于填表方式的关系抽取方法通过为每个关系维护一个表格来表示实体间的关系[14-18]。这些方法将关系抽取任务转化为填写表格的任务，有效地解决了实体嵌套等问题。Ren 等[14]提出了一种面向全局特征的基于填表方式的抽取模型 GRTE，将全局特征反馈到模型中，充分利用了文本的全局特征与局部特征。Wang 等[15]提出了一种统一标签空间的基于表格形式的联合实体关系模型 UniRe，将实体检测看作是关系分类的一种特例。Shang 等[5]提出了在单阶段单模块抽取关系三元组的方法 OneRel，提出了 HornTagging 的表格标记方式，有效解决了实体嵌套的问题。Zhang 等[16]在 OneRel 的基础上引入关系判断模块与边界平滑机制，提出了一种基于特定关系上进行三元组标记与打分的填表方式联合模型 RS-TTS。Ning 等[17]将关系抽取视做在表格中进行目标检测，提出了 OD-RTE 方法。Dai 等[18]提出基于主客体交互和推理路径的表格填充的关系三元组联合抽取的方法 SOIRP。然而，这些方法往往忽略了关系标签本身的语义信息，限制了模型的性能。

1.3 关系标签语义的融合

最近的一些研究关注关系标签的语义信息，并尝试将其融合到关系抽取中[19-21]。Xu 等[19]提出显示引入关系表示的 EmRel，以利用关系、实体和上下文之间丰富的相关性。Cheng 等[20]提出了一种级联双解码器，分别对文本和关系设计了解码器，用以实体和关系的联合抽取。Tang 等[21]提出基于实体与关系统一表示的联合抽取方法 UniRel，在模型中引入关系标签的语义信息，并与文本进行统一建模，解决了文本与关系异构表示的问题。受以上方法的启发，本文将关系抽取建模为计算实体对与关系标签嵌入的相关性分数，将关系标签语义融合到关系抽取中。

1.4 专利文本关系抽取

专利领域文本中蕴含了大量的专业性信息，关系抽取是有效获取其中信息的重要一步。Chen 等[22]建立了英文关系抽取数据集 TFH2020，并提出一种流水线式的专利信息抽取方法。李成奇等[4]建立了一个中文专利实体关系抽取数据集 PERD，并提出最近对寻址的实体关系抽取模型 NPAM，并有效解决了专利文本中常见的实体嵌套问题。邓娜等[23]面向中药专利文本实体关系提出了一种融合语义特征和多层交叉注意力机制的联合抽取模型 TPSCRE。何玉等[24]面向绿色合作专利领域提出了 SpERT-Aggcn，通过注意力引导的图卷积网络引入句法信息，

提高了关系抽取的准确率。王腾科[25]提出基于片段的中文专利实体关系抽取模型SPERE，利用专利文本中重叠关系的语义依赖优化关系抽取模型。这些研究中许多致力于解决重叠关系，为专利文本关系抽取提供了丰富的实践经验，但仍有待改进。本文充分利用关系标签中的语义与文本中的全局信息，进一步提升了专利文本关系抽取的性能。

2 SRL-GFI 模型

SRL-GFI 模型的总体框架如图 1 所示。

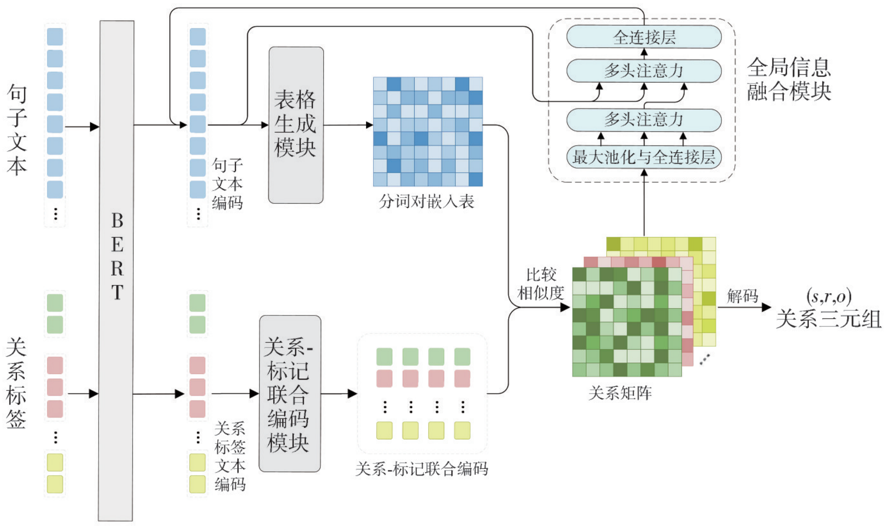

图 1 SRL-GFI 模型的总体架构

Fig. 1 The overall architecture of the SRL-GFI model

首先，利用 BERT 编码器分别对输入的句子和关系标签进行编码，接着计算每个分词对与关系标签的相关性分数。然后通过全局特征融合模块捕捉句子的全局语义信息，并将这些全局特征反馈回模型中，与句子的编码进行融合，再次计算分词对与关系的相关性分数。经若干轮迭代后，对结果进行解码，获得实体关系三元组。整体流程如表 1 所示。

表 1 整体流程

Tab. 1 Overall process

算法 1 SRL-GFI 算法
输入：文本 L，关系集合 R，迭代次数 round 输出：关系三元组集合 *TripleSet*

$H \leftarrow Bert(L)$ //获得句子文本编码
$E_{\text{rel-tag}} \leftarrow RTEmb((Bert(R))$ //通过关系-标记联合编码模块得到关系-标记联合编码
$E_{\text{pair}} \leftarrow TokenPairEmbModule(H)$ //生成分词对嵌入
for i ：= 0 **to** round **do**
1) $T_{\text{score}} \leftarrow CalcScore(E_{\text{pair}}, E_{\text{rel-tag}})$ //计算相关性分数
2) $G \leftarrow GlobalFeatureModule(T_{\text{score}}, H)$ //计算全局特征
3) $E_{\text{pair}} \leftarrow TokenPairEmbModule(G)$ //将全局特征反馈回模型，更新分词对嵌入
end for
$T \leftarrow Softmax(T_{\text{score}})$ //经过 Softmax 函数，得到关系-标记联合标签表
$TripleSet \leftarrow Decoding(T)$ //对关系-标记联合标签表解码，得到关系三元组
return *TripleSet*

2.1 问题描述

给定一段长度为 L 的文本 $S=\{x_1, x_2, \cdots, x_L\}$ 以及 K 个预设的关系 $R=\{r_1, r_2, \cdots, r_K\}$，关系抽取任务的目标是找出文本所蕴含的所有关系三元组 $\{(s_i, r_i, o_i)\}_{i=1}^{N}$，其中 N 是三元组的个数，s_i、o_i 分别是第 i 个三元组的主体、客体，由文本中的连续字符构成，r_i 是主体和客体之间的关系。

2.2 表格标记方式

受 OneRel 的启发，本文采用了 HornTagging 标记方式。该标注方式使用了 4 种标记标签：

1）HB-TB 标注主体的头分词和客体的头分词在表中对应的位置；2）HB-TE 标注主体的头分词和客体的尾分词在表中对应的位置；3）HE-TE 标注主体的尾分词和客体的尾分词在表中对应的位置；4）“-”标注表中上述以外标注的位置。该标记方式可有效解决关系重叠问题。例如在图 2 中，表中记录了“主附件”关系中的标记情况。对于（杀菌装置，主附件，紫外线杀菌灯）、（杀菌装置，主附件，连接头）、（杀菌装置，主附件，定位板）这几个关系三元组，可根据表中标记区别。

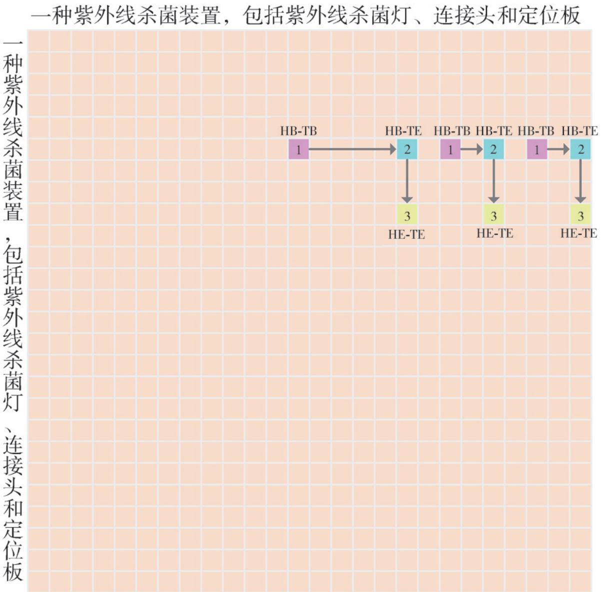

图 2　表格标记示例

Fig. 2　Example for the tagging scheme

2.3　文本编码与分词对嵌入表

对文本 S 通过预训练编码器 BERT 进行编码，得到文本的 d 维隐藏层嵌入 H。

$$H = Bert(S)。\tag{1}$$

然后，通过全连接层，分别生成与主、客体相关的特征 H_s 和 H_o

$$H_s = W_1 H + b_1,\ H_o = W_2 H + b_2,\tag{2}$$

式中：W_1 与 W_2 是可训练的权重；b_1 和 b_2 是可训练的偏置。接着，两者进行哈达玛积运算，得到分词对嵌入表 E_{pair}，嵌入维度为 d。

$$E_{pair} = \{ H_{s,i} \circ H_{o,j} | i,j \leqslant L \}。\tag{3}$$

2.4　关系-标记联合编码

关系标签蕴含了丰富的语义信息，通常以文本形式表示。为充分利用这些语义信息，使用 BERT 预训练模型对关系标签进行编码。关系标签的文本由多个字组成，对关系标签进行分词后的序列长度通常大于 1。为了将这些序列转化为的长度为 1 的序列来作为关系标签的编码，有两种方法可以采用。一种是截取关系标签中的部分字符，如 Tang 等[21]在 UniRel 中，选择最能保留其语义信息的词进行编码。这种方式虽能够融合关系标签文本的语义信息，但存在信息损失，且需要针对不同数据集的关系标签进行人工处理，具有较大的随意性。特别是对于 PERD 数据集的关系标签，找出能较大程度保留其标签语义信息的字符较为困难。另一种是对关系标签文本整体进行编码得到隐藏层序列，然后通过深度学习的方法池化序列，得到关系编码。本文采取后一种方式，使用自注意力机制处理序列以保留标签中各字符的语义信息，并选取其中固定位置的向量作为池化后的关系编码。这种方法考虑到关系标签中每个字符的语义信息，并且可以在训练过程中调整各个字符的权重，从而学习到各字符与关系标签的相关性。

考虑到采用 HornTagging 标记方式，对于每个分词对，模型需要同时判断其对应的关系标签和标记标签。因此，我们设计了关系-标记联合标签，并通过多头注意力机制生成关系-标记联合编码。关系-标记联合编码模块如图 3 所示。

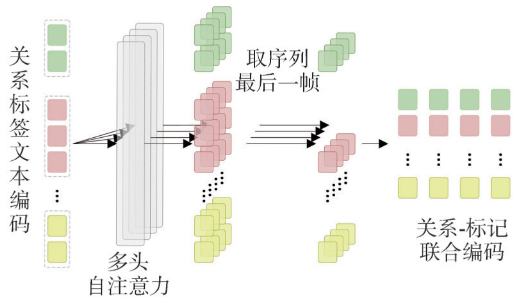

图 3　关系-标记联合编码模块

Fig. 3　Relation-tagging joint encoding module

具体步骤如下：

1）首先，通过 BERT 编码器对关系标签文本 r_i 进行编码，得到相应的隐藏层。

2）然后，利用多头自注意力机制，四个头分别关注四种标记方式（HB-TB、HB-TE、HE-TE、-）的相关信息。

3）最后，对于各关系标签的输出序列，分别取其最后一帧，得到关系-标记联合编码 $E_{rel\text{-}tag}$，如式(4)所示。关系-标记联合编码模块对于每个关系类别和标记方式的组合都进行了编码，编码的嵌入维度为 d。

$$E_{\text{rel-tag}}=\{W_3 MHSA_{t_i}(Bert(r_i))+ b_3|r_i\in R,t_i\in R\},\qquad(4)$$

式中：$MHSA_{t_i}$表示第i个标记方式对应的多头自注意力机制；W_3和b_3为输出的全连接层的权重和偏置。通过BERT预训练模型获取关系标签的语义信息，能够更准确地捕捉关系标签与文本之间的语义关联。

2.5 关系-标记联合判断

关系判断模块对文本分词对嵌入表E_{pair}中每个分词对的嵌入向量同关系-标记联合嵌入$E_{\text{rel-tag}}$计算相关性分数。在Vaswani等[26]提出的Transformer中，使用了一种缩放点积注意力，如式(5)所示。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。\qquad(5)$$

式中：$\boldsymbol{Q}$为查询向量；$\boldsymbol{K}$为键向量；$\boldsymbol{V}$为值向量；d_k为向量的维度。在此过程中，计算查询向量$\boldsymbol{Q}$与键向量$\boldsymbol{K}$的内积得到注意力分数。

借鉴这一方法，将关系-标记联合编码与文本分词对嵌入的相关性分数定义为二者点乘得到的内积。然后加入偏置，得到分词对与关系-标记联合编码的相关性分数表T_{score}。

$$T_{\text{score}}=\{E_{\text{pair }i,j}\cdot E_{\text{rel-tag }r,t}+b_4\},\qquad(6)$$

式中：$E_{\text{pair }i,j}$为分词对(x_i,x_j)的d维嵌入向量；$E_{\text{rel-tag }r,t}$为对第r个关系标签与第t个标记标签的联合编码；b_4为偏置。

2.6 全局特征融合

在专利文本中，文本长度较长，实体之间的关系可能跨越较长的文本距离，且关系三元组中的联系较多。传统的局部特征提取方法往往难以捕捉长距离的语义关联以及关系三元组中的内在联系。为融合全局特征，受GRTE[14]的启发，使用Transformer[26]中的解码器，将分数表中蕴含的全局特征反馈到文本编码与分词对嵌入表模块，进行多次迭代，将最后一次迭代生成的分数表进行解码操作。定义解码层$Decoder(S,S')$

$$Decoder(S,S')=MHA(MHSA(S),S',S'),\qquad(7)$$

式中：MHA表示多头注意力机制；S与S'分别为解码层的原序列和目标序列。

因解码器对全表处理数据量过大，故先使用最大池化和全连接层提取表中与主、客体具有相关性的特征，然后分别同文本原始编码进行解码，得到H_{s}'与H_{o}'

$$H_{\text{s}}'=W_{\text{s}}'Decoder(maxpool_{\text{s}}(T_{\text{score}}),H)+b_{\text{s}}',$$
$$H_{\text{o}}'=W_{\text{o}}'Decoder(maxpool_{\text{o}}(T_{\text{score}}),H)+b_{\text{o}}',\qquad(8)$$

式中：$maxpool_{\text{s}}$与$maxpool_{\text{o}}$这两种最大池化操作分别用来突出与主、客体相关的特征；W_{s}'与W_{o}'是可训练的权重，b_{s}'与b_{o}'是可训练的偏置。

接着使用残差的方式分别与当前的主、客体特征H_{s}和H_{o}进行融合，得到新的主、客体特征$H_{\text{s}}^{\text{next}}$和$H_{\text{o}}^{\text{next}}$。

$$H_{\text{s}}^{\text{next}}=H_{\text{s}}+H_{\text{s}}',\ H_{\text{o}}^{\text{next}}=H_{\text{o}}+H_{\text{o}}'。\qquad(9)$$

最后，将新的主、客体特征反馈回分词对嵌入表生成模块，进行下一次迭代。通过引入全局特征融合模块，模型能够更好地捕捉文本中的长距离依赖关系以及关系三元组之间的内在联系，从而提升关系抽取的准确性。

2.7 解　码

经过若干次迭代后，使用Softmax函数处理分词对与关系-标记联合编码的相关性分数表T_{score}，得到关系-标记联合标签表T，这个表记录了头、尾实体的边界及其对应的关系。接着根据HornTagging的标记方式，对得到的表格进行解码，以获取句子中蕴含的实体关系三元组。在各关系对应的分表中，从每个“HB-TB”出发，找到其对应的“HB-TE”作为三元组的主体，再从“HB-TE”出发找到“HE-TE”作为三元组的客体。在存在实体重叠的情况下，优先选取靠近“HB-TB”最短的实体。如图2所示，按照箭头方向找出相应的分词对进行解码。对于(杀菌装置，主附件，紫外线杀菌灯)这个关系三元组，先找到标记为HB-TB的分词对(杀，紫)，再找出标记为HB-TE的分词对(杀，灯)，接着找到标记为HE-TE的分词对(置，灯)，得到三元组(杀菌装置，主附件，紫外线杀菌灯)。

2.8 损失函数

采用交叉熵损失函数来衡量预测分数与真实标签之间的差异，并将其作为模型训练过程中的优化目标。损失函数为

$$\mathcal{L}=-\frac{1}{L\times L\times K}\times \sum_{i=1}^{L}\sum_{j=1}^{L}\sum_{k=1}^{K}\log P(y_{(x_i,x_j,r_k)}=g_{(x_i,x_j,r_k)}|S),\qquad(10)$$

式中：$g_{(x_i,x_j,r_k)}$表示真实的关系-标记联合标签。

3 实 验

3.1 数据集

实验在多个数据集上进行，包括PERD[4]、TFH2020[22]、NYT[27]、CONLL04[28]与SCIERC[29]数据集。PERD为李成奇等[4]提出的中文专利数据集，TFH2020为Chen等[22]提出的面向硬盘薄膜磁头技术的英文专利数据集，本文以8∶1∶1的比例将其划分为训练集、验证集和测试集，其余均为英文公开数据集。为方便实验，本文将CONLL04与SCIERC处理为与NYT相同的格式。各数据集的统计数据如表2所示。

表2 数据集统计数据

Tab. 2 Dataset statistics

数据集	PERD	TFH2020	NYT	CONLL04	SCIERC
关系类型数	8	15	24	5	7
句子总数	4 010	1 010	66 196	1 441	2 687
三元组均数	12.52	18.09	1.58	1.42	1.73
训练集	3 210	808	56 196	922	1 861
验证集	400	101	5 000	231	275
测试集	400	101	5 000	288	551

3.2 评价指标

本文采用精确率(R_{prec})、召回率(R_{rec})和$F1$值($F1$)作为评价指标计算关系三元组抽取结果，验证模型的有效性，公式定义如式(11)～式(13)所示。对于一个实体关系三元组，评判标准为三元组中的实体对及其关系相互匹配，则视为正确。本文对实体采用完全匹配的模式，即在实体的整个跨度都预测成功时，才视为正确。

$$R_{prec}=\frac{N_{TP}}{N_{TP}+N_{FP}}, \tag{11}$$

$$R_{rec}=\frac{N_{TP}}{N_{TP}+N_{FN}}, \tag{12}$$

$$F1=2\times\frac{R_{prec}\times R_{rec}}{R_{prec}+R_{rec}}, \tag{13}$$

式中：N_{TP}是被正确预测的正样本的个数；N_{FP}是实际为负样本被错误预测为正样本的个数；N_{FN}是实际为正样本被错误预测为负样本的个数。

3.3 实验设置

本文以Python 3.8.0为开发语言，Pytorch 1.8.0作为开发框架，在一台显卡配置为NVIDIA GeForce RTX 3090的服务器上进行实验，操作系统为Ubuntu 20.04。对于不同的数据集，采用各自合适的BERT预训练模型，词嵌入维度d均为768维。对于PERD中文专利数据集，预训练模型采用“Chinese-bert-wwm-ext”；对于NYT和CONLL04数据集，采用“bert-base-cased”；对于SCIERC数据集，采用“scibert-scivocab-uncased”。文本最大长度*max_len*对于PERD与TFH2020两个专利数据集设置为300，其余均设置为100。批量大小*batch_size*对于PERD与TFH2020两个专利数据集设置为4，其余设置为8。全局特征融合模块的迭代次数在TFH2020数据集上为2，在其余数据集上为1。另外，由于Horntagging标记方式难以处理单个分词组成的实体，在NYT、CONLL04、SCIERC数据集上分词中间添加“[unused1]”以便抽取单分词实体。在PERD数据集上由于单分词构成的实体极少，而文本长度长，不做添加。在TFH2020数据集上，由于其中单分词实体比例较大，且文本长度长，采用英文单词中间而非分词中间添加“[unused1]”的分词方式解决上述问题。优化器选取为Adam来最小化损失函数，学习率设置1×10^{-5}，训练阶段使用早停策略以防止过拟合。

3.4 实验结果

本文选择了一些已发表的方法作为基线模型进行对比实验。

1) CasRel[10]：一种基于层叠指针网络的关系抽取方法，将关系建模为从主语到宾语的函数；

2) RIFRE[30]：一种基于异构图神经网络的表示迭代融合的关系抽取方法；

3) TPLinker[13]：一种基于握手标记方案的端到端序列标注关系抽取模型；

4) OneRel[5]：一种在单阶段通过单个模块的关系抽取的SOTA方法，提出一种HornTagging标记方式解决实体重叠问题，缓解了级联误差传播以及信息冗余的问题；

5) NPAM[4]：一种最近对寻址的实体关系抽取方法。

实验结果如表3所示，展示了不同模型的性能对比。表中标有*的为引用其他论文的结果，其他为复现结果；加粗部分表示在特定指标上获得的最高分数。SRL-GFI在各数据集上准确率、召回率、$F1$值均超过基准模型OneRel，其中在PERD数据集上的准确率、召回率、$F1$值三个指标分别提升了5.2,

4.8,5.0百分点,在TFH2020数据集上分别提升了3.1百分点,0.8百分点,1.8百分点,在CONLL04数据集上分别提升了8.4百分点,1.2百分点,4.8百分点,在SCIERC上分别提升了1.2百分点,4.3百分点,2.7百分点,在NYT数据集上准确率、$F1$值分别提升了1.3百分点,0.5百分点,召回率降低了0.1百分点。

与其他的基线模型对比,该模型表现出相当的优势,在各数据集上$F1$值均达到了最大值。其中,在PERD上达到79.8%,在TFH2020上达到了55.9%,在NYT上达到92.6%,在CONLL04上达到67.4%,在SCIERC上达到44.2%。

表3 对比实验结果

Tab.3 Result of comparative experiment %

模型	PERD			TFH2020			NYT			CONLL04			SCIERC		
	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$
CasRel	68.8*	51.6*	59.0*	45.6	33.1	38.8	89.7*	89.5*	89.6*	61.4	59.5	60.4	35.8	31.8	33.7
RIFRE	66.0*	68.2*	67.1*	53.0	35.3	42.4	**93.6***	90.5*	92.0*	65.1	**69.7**	67.3	35.9	30.3	32.9
TPLinker	58.2*	69.9*	63.5*	45.6	36.6	40.6	91.4*	**92.6***	92.0*	56.2	63.9	59.8	**50.2**	37.6	43.0
OneRel	73.4	76.3	74.8	58.9	50.1	54.1	91.7	92.4	92.1	60.4	64.9	62.6	43.0	40.0	41.5
NPAM	75.6*	70.1*	72.7*	49.6	36.7	42.2	91.5	90.3	90.9	60.8	58.2	59.5	30.0	33.8	31.8
SRL-GFI	**78.6**	**81.1**	**79.8**	**62.0**	**50.9**	**55.9**	93.0	92.3	**92.6**	**68.8**	66.1	**67.4**	44.2	**44.3**	**44.2**

实验结果反映了模型的有效性。SRL-GFI首先通过挖掘关系的语义信息,生成关系-标记联合编码,使得模型能够更准确地捕捉关系标签与文本之间的语义关联。其次,模型融合文本的全局特征,将文本中的全局语义信息反馈回模型中,进一步优化了实体关系抽取,增强了对复杂文本的理解能力。SRL-GFI解决了传统模型中对关系与文本编码、文本间的内在联系利用不充分的问题,提升了模型的性能,使其能够高效地对专利领域的复杂文本进行关系抽取。

3.5 消融实验

为探究模型的关系编码模块以及全局特征融合模块对于模型性能的影响,进行了消融实验,结果如表4所示,其中包含两方面的消融实验:

1)针对关系编码模块的消融实验,如表中de-RelEmb所示。将关系标签编码模块替换为全连接层,以消除关系标签语义对模型的影响。

2)针对全局特征融合模块的消融实验,如表中de-GFM所示。将全局特征融合模块的迭代次数设置为0,不再计算文本的全局特征,消除文本全局特征对模型的影响。

从表中可以观察到,完整的模型SRL-GFI总体性能最佳,去掉或者替换任意一个模块都会导致一定程度的性能下降,这说明关系标签语义与全局特征融合对于模型的性能均有提升作用。其中关系标签语义增强了对实体对的关系判断,对模型性能的提升起到主要的作用;全局特征融合利用了关系之间的联系与句子文本中的全局特征,可更好地表示句子文本信息,该模块对模型性能的提升也起到积极作用,但较关系编码模块效果弱。

表4 消融实验结果

Tab.4 Result of ablation experiment %

模型	PERD			TFH2020			NYT			CONLL04			SCIERC		
	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$
de-RelEmb	79.0	78.3	78.7	60.5	50.5	55.1	92.7	91.6	92.1	61.4	64.1	62.7	**46.0**	41.1	43.4
de-GFM	78.3	80.0	79.1	60.9	49.9	54.9	92.5	92.1	92.3	68.1	**66.1**	67.1	44.9	43.3	44.1
SRL-GFI	**78.6**	**81.1**	**79.8**	**62.0**	**50.9**	**55.9**	**93.0**	**92.3**	**92.6**	**68.8**	**66.1**	**67.4**	44.2	**44.3**	**44.2**

全局特征融合模块GFM在其他文本中的表现不如在专利文本中,原因是其句子长度较短,单个句子中的关系三元组不多,GFM的作用不如在专利数据集中明显。且在模型中有了关系标签语义模块后,关系标签与文本的关系被考虑进来,GFM的效果就不凸显了。这表明,GFM在处理复杂文本时具有一定的优势,但在简单文本中,其作用可能被其他模块所替代。此外,GFM会增加计算量,导致模型的训练时间增加。

3.6 参数分析实验

为分析参数对模型的影响,在PERD和TFH2020数据集上进行参数分析实验,重点关注全局信息融合模块的迭代数对模型性能的影响,

实验结果如表5所示。

表5 参数分析实验

Tab. 5 Parameter analysis experiment %

迭代次数	PERD			TFH2020		
	R_{prec}	R_{rec}	$F1$	R_{prec}	R_{rec}	$F1$
0	78.3	80.0	79.1	60.9	49.9	54.9
1	78.6	81.1	**79.8**	61.9	50.6	55.7
2	79.1	80.1	79.6	62.0	**50.9**	**55.9**
3	**79.5**	79.0	79.2	**62.5**	50.5	55.8
4	77.2	**81.7**	79.4	61.1	50.6	55.6

实验结果表明，在PERD数据集上，一次全局信息融合迭代即可有效融合全局特征并达到最优性能，表明模型的感受野已经足够覆盖全局信息，并且足以利用关系三元组之间的关系。而在TFH2020数据集上变化比较平稳，由于其全局信息更为复杂，单个句子中三元组数目更多，在迭代次数为2时其性能才达到最大。

3.7 细粒度实验

为进一步探究模型在中文专利数据集上的作用，分别按照句子文本中的关系三元组数目、关系重叠类型和关系类别在PERD数据集上进行实验，实验结果如图4～图6所示。

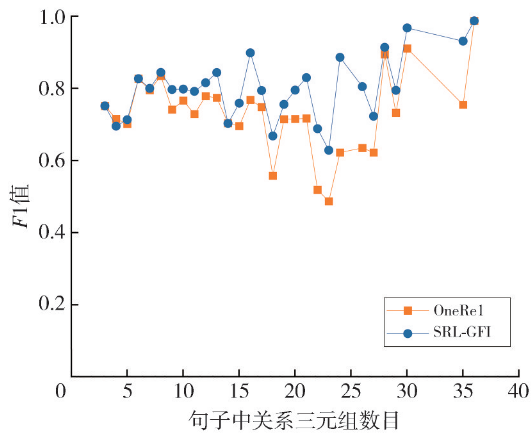

图4 不同关系三元组数目的实验结果

Fig. 4 Results on different triple numbers

图4展示了不同关系三元组数目的实验结果，可以观察到模型对各关系三元组数目的句子的抽取效果大多有所提升。图5展示了不同类型的关系重叠情况下的实验结果，表明在EPO、SEO与SOO等不同情况下的抽取效果均有所提升，说明SRL-GFI在中文专利文本领域的关系抽取任务可有效解决关系重叠问题。图6展示了不同关系类别的实验结果，结果表明在对各关系类别的关系抽取均有性能提升。SRL-GFI模型在中文专利文本关系抽取任务的各方面都表现出了显著的性能提升。

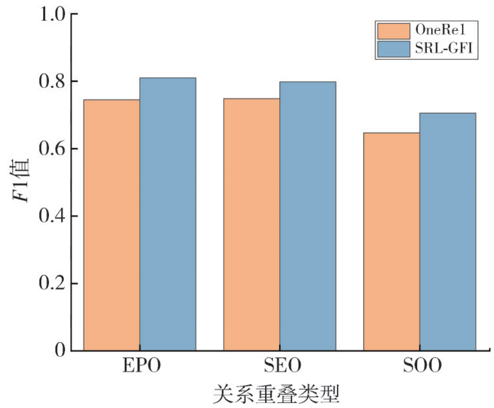

图5 不同关系重叠类型的实验结果

Fig. 5 Results on different overlapping patterns

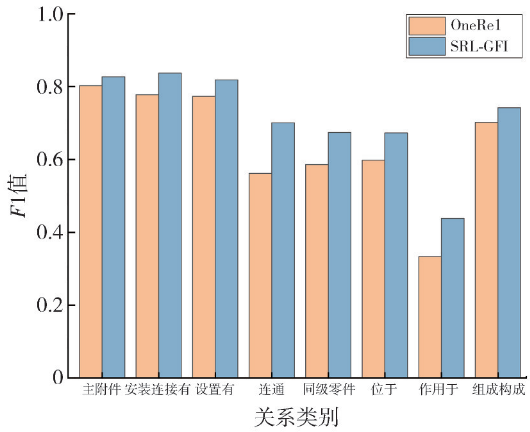

图6 不同关系类别的实验结果

Fig. 6 Results on different relations

4 结论与展望

本文介绍了一种基于关系标签语义与全局信息融合的关系抽取方法SRL-GFI，针对专利数据集中重叠关系数目多的特点，选取OneRel作为基准模型进行关系抽取，并在此基础上针对OneRel关系标签语义信息缺失以及文本的全局特征融合不足的问题进行优化，取得了性能提升。SRL-GFI模型在PERD、TFH2020两个专利文本数据集上以及NYT、CONLL04、SCIERC通用数据集上的性能也均有提升，且与其他的现有的方法相比具有相当的竞争力，展示了所提方法优良的泛化性能。另外，通过对SRL-GFI的关键模块的验证，证实了本文所提的关系标签语义与全局特征融合对于关系抽取任务的有效性。未来工作将探索结合文本和关系标签的BERT编码特征，以进一步优化实体对与关系标签相关性分数的计算方

法。此外,考虑到一些数据集上的关系类别分布不平衡的问题,以及HornTagging标记方式所带来的关系-标记联合标签不平衡的问题,后续可以对标签分布的问题进行研究,优化标记方式及损失函数,提高关系抽取模型的性能。

参考文献:

[1] 陈亮,陈利利,许海云,等.国内外专利挖掘研究进展与前瞻[J].图书情报工作,2024,68(2):110-133.
CHEN Liang, CHEN Lili, XU Haiyun, et al. A global literature review in recent advancement of patent mining[J]. Library and Information Service, 2024, 68(2): 110-133. (in Chinese)

[2] GOLSHAN P N, DASHTI H R, AZIZI S, et al. A study of recent contributions on information extraction [DB/OL]. (2018-05-15) [2024-11-21]. https://arxiv.org/abs/1803.05667v1.

[3] 刘烨宸,李华昱.领域知识图谱研究综述[J].计算机系统应用,2020,29(6):1-12.
LIU Yechen, LI Huayu. Survey on domain knowledge graph research[J]. Computer Systems & Applications, 2020, 29(6): 1-12. (in Chinese)

[4] 李成奇,雷海卫,李帆,等.最近对寻址的专利实体关系抽取方法[J].计算机工程与设计,2024,45(4):1100-1108.
LI Chengqi, LEI Haiwei, LI Fan, et al. Method for extracting patent entity relationships based on nearest pairing addressing [J]. Computer Engineering and Design, 2024, 45(4): 1100-1108. (in Chinese)

[5] SHANG Y M, HUANG H, MAO X. Onerel: Joint entity and relation extraction with one module in one step[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 11285-11293.

[6] 鄂海红,张文静,肖思琪,等.深度学习实体关系抽取研究综述[J].软件学报,2019,30(6):1793-1818.
E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning[J]. Journal of Software, 2019, 30(6): 1793-1818. (in Chinese)

[7] CHAN Y S, ROTH D. Exploiting syntactico-semantic structures for relation extraction [C]//49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011: 551-560.

[8] GORMLEY M R, YU M, DREDZE M. Improved relation extraction with feature-rich compositional embedding models[C]//2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1774-1784.

[9] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//52nd Annual Meeting of the Association for Computational Linguistics, 2014: 402-412.

[10] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction [C]//58th Annual Meeting of the Association for Computational Linguistics, 2020: 1476-1488.

[11] YAN Z, ZHANG C, FU J, et al. A partition filter network for joint entity and relation extraction[C]//2021 Conference on Empirical Methods in Natural Language Processing, 2021: 185-197.

[12] ZHENG H, WEN R, CHEN X, et al. PRGC: Potential relation and global correspondence based joint relational triple extraction[C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 6225-6235.

[13] WANG Y, YU B, ZHANG Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking[C]//28th International Conference on Computational Linguistics, 2020: 1572-1582.

[14] REN F, ZHANG L, YIN S, et al. A novel global feature-oriented relational triple extraction model based on table filling [C]//2021 Conference on Empirical Methods in Natural Language Processing, 2021: 2646-2656.

[15] WANG Y, SUN C, WU Y, et al. UniRE: A unified label space for entity relation extraction [C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 220-231.

[16] ZHANG J, JIANG X, SUN Y, et al. RS-TTS: A novel joint entity and relation extraction model[C]//26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2023: 71-76.

[17] NING J, YANG Z, SUN Y, et al. OD-RTE: A one-stage object detection framework for relational triple extraction[C]//61st Annual Meeting of the Association for Computational Linguistics, 2023: 11120-11135.

[18] DAI Q, YANG W, WANG L, et al. SOIRP:

Subject-object interaction and reasoning path based joint relational triple extraction by table filling[J]. Neurocomputing, 2024, 580: 127492.

[19] XU B, WANG Q, LYU Y, et al. EmRel: Joint representation of entities and embedded relations for multi-triple extraction[C]//2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 659-665.

[20] CHENG J, ZHANG T, ZHANG S, et al. A cascade dual-decoder model for joint entity and relation extraction [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9: 1130-1142.

[21] TANG W, XU B, ZHAO Y, et al. UniRel: Unified representation and interaction for joint relational triple extraction[C]//2022 Conference on Empirical Methods in Natural Language Processing, 2022: 7087-7099.

[22] CHEN L, XU S, ZHU L, et al. A deep learning based method for extracting semantic information from patent documents[J]. Scientometrics, 2020, 125(1): 289-312.

[23] 邓娜，喻卓群，但文俊，等. 一种融合语义特征和多层交叉注意力机制的中药专利文本实体关系联合抽取模型[J]. 数据分析与知识发现，2025, 9(7): 141-153.
DENG Na, YU Zhuoqun, DAN Wenjun, et al. A joint extraction model for entity relationship in traditional Chinese medicine patent texts based on semantic features and multi-layer cross-attention mechanism[J]. Data Analysis and Knowledge Discovery, 2025, 9(7): 141-153. (in Chinese)

[24] 何玉，张晓冬，郑鑫. 基于 SpERT-Aggcn 模型的专利知识图谱构建研究[J]. 数据分析与知识发现，2024, 8(1): 146-156.
HE Yu, ZHANG Xiaodong, ZHENG Xin. Constructing patent knowledge graph with spert-aggcn model [J]. Data Analysis and Knowledge Discovery, 2024, 8(1): 146-156. (in Chinese)

[25] 王腾科. 面向中文专利文本的命名实体识别及关系抽取研究[D]. 淮南：安徽理工大学，2024.

[26] VASWANI A, SHAZEER N, PARMAR N. et al. Attention is all you need [DB/OL]. (2017-06-12) [2024-11-21]. https://arxiv.org/abs/1706.03762.

[27] RIEDEL S, YAO L, MCCALLUM A. Modeling relations and their mentions without labeled text[C]//European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010), 2010: 148-163.

[28] ROTH D, YIH W. A linear programming formulation for global inference in natural language tasks[C]//8th Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, 2004: 1-8.

[29] LUAN Y, HE L, OSTENDORF M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[C]//2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3219-3232.

[30] ZHAO K, XU H, CHENG Y, et al. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction [J]. Knowledge-Based Systems, 2021, 219: 106888.